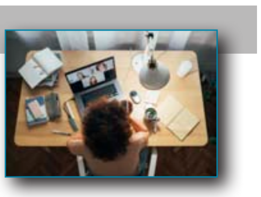

Genetic Risk Scores for Breast Cancer Based on Machine Learning

BY WARREN FROELICH

At birth, it's recently been observed that the length of our germline chromosomes varies from individual to individual, as unique as a fingerprint or our DNA. Now, researchers from the University of California Irvine have found they can teach a computer through machine learning to predict a woman's risk of developing breast cancer in her lifetime by analyzing variations in chromosomal length.

Though in its early stages, the new technique has yielded some impressive results with breast cancer and other tumors that, according to the researchers, could soon become a standard test.

"Genetic tests that predict whether or not you will develop cancer at some point in your life are coming soon," said James P. Brody, PhD, Associate Professor of Bioengineering at UC Irvine, who presented early results at the AACR Virtual Special Conference: Artificial Intelligence, Diagnosis, and Imaging. "These tests are getting much more accurate. In the next decade or so, these tests could be in widespread use."

As outlined during his talk, Brody noted that current predictors for cancer risk were largely based on the additive effects of single nucleotide polymorphisms in germline DNA. An example includes tests that identify BRCA1 and BRCA2 to determine potential risk for breast cancer.

"These tests work really for a subset of people, but are meaningless for most people," he noted. But they also fail to account for epistatic interactions, he said, where the actions of one gene are affected by the presence of other genes.

Machine learning, a form of artificial intelligence in which a model automatically learns and improves based on previous experiences, can detect these associations. But even this hasn't done particularly well as a predictor of disease, owing to the overwhelming number of features needed to be analyzed—maybe 10,000 different patients with a million different variables.

What to do? "Try to increase the number of patients," said Brody. "That's expensive and time-consuming. Or you can somehow try to decrease the number of entries, somehow condense information in the genome. That's what we're trying to do, and the number we're working on is chromosomal length."

Study Details

In his UC Irvine lab, Brody and colleagues sought to use modern machine learning algorithms to identify complex patterns in germline DNA. They hit on something unique that might work: variations in germline chromosomal length.

As outlined in his talk, Brody said germline chromosomes vary from person to person—like a genetic signature—generally resulting from many different structural variants, such as deletions, translocations, and duplications of DNA segments, also collectively referred to as copy number variations.

"We can take this problem, which…had a million different things going across here, and reduce it to like 22 different numbers (representing the length of each chromosome in germline DNA, aside from the sex chromosomes)," Brody explained. "Now we have a good system set up for machine learning."

For their model, Brody and team employed the H2O platform for machine learning to test four different algorithms, with training done on a desktop computer. The algorithms in question included a generalized linear model, distributed random forest, gradient boosting machine, and deep learning model.

Using these algorithms, the computer was taught to distinguish between patients diagnosed with one cancer (i.e., breast cancer) and those not diagnosed with breast cancer.

"We are using the same technology and methods that Google/Facebook use to determine whether a collection of pixels depicts a cat," explained Brody in an interview. "Instead of pixels, we have measurements on germline DNA. Instead of a cat, we determine if the person had a particular form of cancer."

iStock

Initially, the study used data from The Cancer Genome Atlas (TCGA), sponsored by the National Institutes of Health, a large project that characterizes molecular differences in 33 different human cancers. These include tissue samples from tumors, normal tissues adjacent to the tumor, and normal blood samples.

From the TCGA database, the team extracted chromosomal length variations from peripheral blood samples of 968 cases of female breast cancer, along with 3,715 controls—all women who never had breast cancer. Diagnosis for each of the cases in the TCGA database was confirmed by a pathologist.

Results showed their machine learning model was able to distinguish women with breast cancer from those with no history of the disease with an AUC (area under the curve) of about .75.

"AUCs are easy to interpret," Brody said. "They are the accuracy we would get if we did 1,000 tests, where 500 were known to be true and 500 were known to be false."

Just like a true/false test you would take in school, it's easy to get 0.50 by just guessing, so a 0.50 grade is virtually useless. A perfect score is 1.00.

An AUC is a statistic that is regularly used for diagnostic tests. For example, a mammogram has an AUC of about 0.95 for diagnosing breast cancer. Predictive tests generally have lower AUCs than diagnostic tests. For instance, a predictive test for coronary artery disease using genetic information has an AUC of .62 and the best predictive test now available for cancer has an AUC of .69. So, the initial results from the TCGA dataset caught Brody by surprise.

"I was surprised and excited that night," he said. "The next day I was worried." Brody was concerned because there are technical errors in machine learning that will lead to "amazing-looking results." He spent several months trying to build confidence in the results, and then he began worrying he may have uncovered technical problems in the way the data was generated.

To alleviate his concerns, Brody tried to see if he could replicate his results with a completely different set of data. Here, he turned to the UK Biobank, consisting of half a million people ages 40-69. From this population, Brody and team identified 1,534 cases who both self-reported and were identified by cancer registries as having a diagnosis of breast cancer. Some 4,391 women with no history of breast cancer served as a control group.

The genetics for each of the women were quantified with 88 numbers, each representing the length variation of one-quarter of each of the 22 chromosomes analyzed. The X chromosome was not used. Results showed Brody's machine learning model for breast cancer risk, using data from the UK Biobank, had an AUC of about .81. "When I saw the results on that dataset, I was satisfied," he said. As for next steps, Brody said he would like to see the prognostic value of his machine learning model improve for breast cancer before taking the technology into the clinic. Right now, he'd give his present model a grade of about a B-. "It's much better than others, but not great," he said.

Brody's also creating models for other cancers, including one for ovarian cancer (soon-to-be-published). "The method should be generally applicable to predict any cancers or other complex genetic conditions, including conditions as diverse as heart disease and schizophrenia," he noted. OT

Warren Froelich is a contributing writer.